Proceedings of the 18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics (CIBB 2023)

GEN-RWD Sandbox: A Modular Architecture for Privacy-preserving Data Sharing for AI-driven Medical Research

Benedetta Gottardelli[1,*], Leonardo Nucciarelli[1], Erica Tavazzi[2], and Andrea Damiani[3]

[1] Department of Diagnostic Imaging, Radiation Oncology and Hematology,Università Cattolica del Sacro Cuore, Rome, Italy
[2] Department of Information Engineering, University of Padova, Padova, Italy
[3] Gemelli Generator Real World Data, Fondazione Policlinico Universitario Agostino Gemelli (IRCCS), Rome, Italy

*corresponding author: benedetta.gottardelli@gmail.com

Keywords: *privacy-preserving data sharing; personalised medicine; distributed analytics.*

Abstract. AI-driven research has become crucial in healthcare, relying on vast amounts of data for patient-specific treatments. However, sharing hospital data is challenging due to privacy concerns and legal restrictions. We propose the Gemelli Generator - Real World Data (GEN-RWD) Sandbox, a modular architecture that provides a secure environment for external stakeholders such as researchers, pharmaceutical companies, and policy makers to interact with hospital data. The GEN-RWD Sandbox consists of two modules within the hospital premises and an internet-based web interface for external users. This architecture ensures data privacy and security while enabling users to select cohorts of interest, define inclusion criteria, and perform analyses through the user-friendly web interface. Its modular design allows scalability and adaptability to accommodate multiple centers in a distributed analytics environment, facilitating broader research studies.

1 Introduction

Artificial intelligence techniques have been used to enhance patient conditions and treatments, leading to an increasing need of data availability [1]. However, due to the sensitivity of patients' personal information, data privacy is a big concern in medical research. GDPR [2] was drafted to balance the demand for security and privacy and the demand for access to data. Federated learning, or Distributed Analytics, [3] has gained popularity and application as it allows data to remain localized and distributed while conducting analysis, minimizing the risk of data breaches. It respects data ownership, allowing entities to retain control over their data while participating in collaborative research. In certain cases, federated learning is preferred over computationally intensive techniques such as homomorphic encryption [4] and secure multiparty computation [5]. Many privacy-preserving infrastructures based on the federated learning principles have been proposed to fulfill different use cases in healthcare domain [6, 7, 8, 9, 10, 11, 12, 13]. This paper introduces the Generator - Real World Data (GEN-RWD) Sandbox, a proposal for a distributed analysis platform that serves as a playground for non-technical researchers with limited programming knowledge. It enables them to analyze clinical data without the need for data transfer from the hosting institution. The name ”sandbox” derives from the provided secure and confined environment, where users can explore and experiment with the data without gaining possession of it. In Section 2 we present our architecture for the GEN-RWD Sandbox and provide a detailed description of its modules. In Section 3 we outline the characteristics of the experimental setting. In Section 4 we discuss our solution and we give an outlook for future work.

2 Architecture

The purpose of the GEN-RWD Sandbox is to allow external non-technical users to perform analyses on hospital data without the need for data sharing. The underlying principle is that authorized external users submit their research requests through a GUI, which are then processed within the hospital. Only the results, from which it is not possible to trace back to patient-level original data, are shared with the external users outside the hospital's domain. The GEN-RWD Sandbox is designed as a modular architecture consisting of three main modules: GUI, Proxy, and Processor (Figure 1). Each module has a specific role within the system to facilitate efficient interaction and analysis execution.

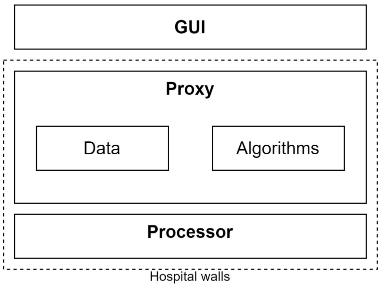


Figure 1: GEN-RWD Sandbox's modular architecture consisting of three main modules: GUI, Proxy, and Processor

2.1 Processor

The Processor module is the central component of the infrastructure, responsible for executing tasks according to a black box model that relies on a minimal structure (Figure 2). This structure comprises a token, that contains the task details ad accessory files, an input folder, the processing unit, and an output folder. The Processor initiates the execution of a task when a token is inserted in its input folder (Figure 2 **a**). While processed, the token is removed from the input folder thus allowing any other accumulated tokens to be evaded when the processing is finished. Upon completion of the computation, the Processor generates an output token, which is deposited into the output folder (Figure 2 **b**).The Processor module is designed as a flexible task executor. It is indeed able to run R not only and Python scripts but also Docker images. Moreover, the Processor is able to process tasks in parallel launching each execution in background and keeping track of their respective logs.

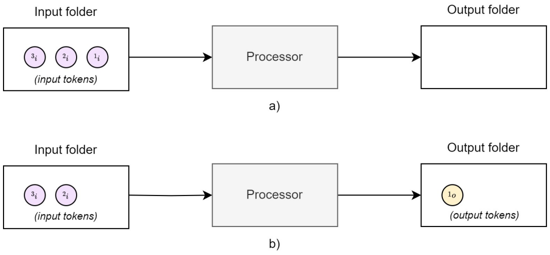


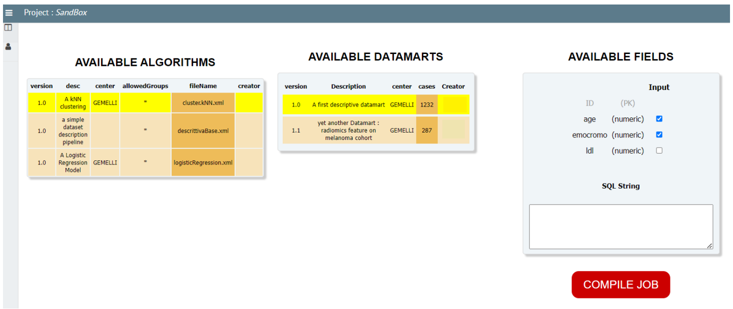
Figure 2: The fundamental architecture of the Processor module.

2.2 Proxy

The Proxy module is responsible for managing the interaction between the GUI and the computing unit, i.e. the Processor. It handles the function of storing and consolidating data and algorithms. It has three additional main functions: monitoring the GUI for new tasks to be executed and preparing the tokens for the Processor module, forwarding logs produced by the Processor to the GUI, and monitoring the output folder of the Processor for output results to send to the GUI. The Proxy module communicates with the Processor via the filesystem, while communications with the GUI are handled in pooling via https on port 80. The modularity of the architecture allows the Proxy module to forward requests received from the higher layer (GUI) to multiple Processors, while also assuming the role of a dispatcher for computational load. At the same time, the Proxy module can be configured to receive instructions from multiple GUIs.

2.3 Graphical User Interface (GUI)

The GEN-RWD Sandbox framework uses a GUI to enable users with a basic understanding of data analysis and machine learning to choose objects from a catalog and submit analysis requests. The GUI allows users to access a list of available algorithms and datamarts and to compose the desired analytic task via interactive selection (Figure 3). Additionally, the GUI features a JOB Status section that displays a list of sent jobs and the results are accessible as HTML reports (Figure 4). When a job is submitted by the user, the GUI includes all the details within an XML file and proposes it to a specific URL, which is continuously monitored by the associated instances of the Proxy. The GUI can be associated with multiple proxies, offering an algorithm and data catalog that extends across multiple hospitals.



AVAILABLE ALGORITHMS

version	desc	center	allowedGroups	fileName	creator
1.0	A kNN clustering	GEMELLI	*	cluster.kNN.xml	
1.0	a simple dataset description pipeline		*	descrittivaBase.xml	
1.0	A Logistic Regression Model	GEMELLI	*	logisticRegression.xml	

AVAILABLE DATAMARTS

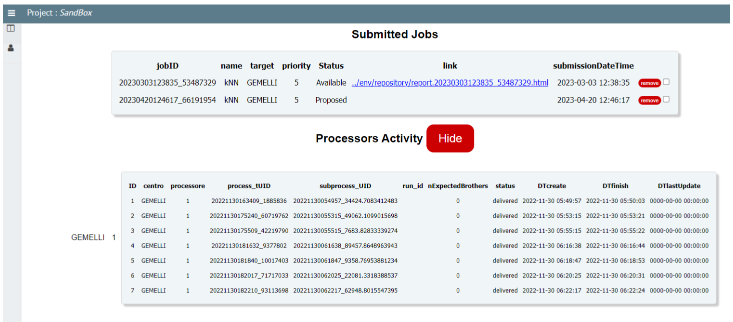
version	Description	center	cases	Creator
1.0	A first descriptive datamart	GEMELLI	1232	
1.1	yet another Datamart : radiomics feature on melanoma cohort	GEMELLI	287	

AVAILABLE FIELDS

		Input
ID	(PK)	
age	(numeric)	☑
emocromo	(numeric)	☑
ldl	(numeric)	☐

SQL String

Figure 3: GEN-RWD Sandbox GUI's *Job submission* page.



Submitted Jobs

jobID	name	target	priority	Status	link	submissionDateTime		
20230303123835_53487329	kNN	GEMELLI	5	Available	../env/repository/report.20230303123835_53487329.html	2023-03-03 12:38:35	remove	☐
20230420124617_66191954	kNN	GEMELLI	5	Proposed		2023-04-20 12:46:17	remove	☐

Processors Activity

GEMELLI 1

ID	centro	processore	process_tUID	subprocess_UID	run_id	nExpectedBrothers	status	DTcreate	DTfinish	DTlastUpdate
1	GEMELLI	1	20221130163409_1885836	20221130054957_34424.7083412483		0	delivered	2022-11-30 05:49:57	2022-11-30 05:50:03	0000-00-00 00:00:00
2	GEMELLI	1	20221130175240_60719762	20221130055315_49062.1099015698		0	delivered	2022-11-30 05:53:15	2022-11-30 05:53:21	0000-00-00 00:00:00
3	GEMELLI	1	20221130175509_42219790	20221130055515_7683.82833339274		0	delivered	2022-11-30 05:55:15	2022-11-30 05:55:22	0000-00-00 00:00:00
4	GEMELLI	1	20221130181632_9377802	20221130061638_89457.8648963943		0	delivered	2022-11-30 06:16:38	2022-11-30 06:16:44	0000-00-00 00:00:00
5	GEMELLI	1	20221130181840_10017403	20221130061847_9358.76953881234		0	delivered	2022-11-30 06:18:47	2022-11-30 06:18:53	0000-00-00 00:00:00
6	GEMELLI	1	20221130182017_71717033	20221130062025_22081.3318388537		0	delivered	2022-11-30 06:20:25	2022-11-30 06:20:31	0000-00-00 00:00:00
7	GEMELLI	1	20221130182210_93113698	20221130062217_62948.8015547395		0	delivered	2022-11-30 06:22:17	2022-11-30 06:22:24	0000-00-00 00:00:00

Figure 4: GEN-RWD Sandbox GUI's *Job status* page.

2.4 Communication flow

Communication between the three modules of GEN-RWD Sandbox occurs asynchronously, with the Proxy module being responsible for managing communication with the upper level, which is the GUI, and the lower level, which is the Processor. The proxy is always listening for new submissions from the GUI and constantly monitoring the processor's folders for new logs or results. In turn, the Processor continuously monitors its input folder to detect any instructions forwarded by the Proxy. Being on the same computer network within the hospital premises, the Proxy and the Processor communicate asynchronously through the filesystem. However, since the GUI is located on the internet outside the hospital's computer perimeter, the proxy communicates with it through pool requests via HTTPS on port 80.

Figure 5 depicts an example of a message exchange that occurs when a job is submitted by the user within the GUI.

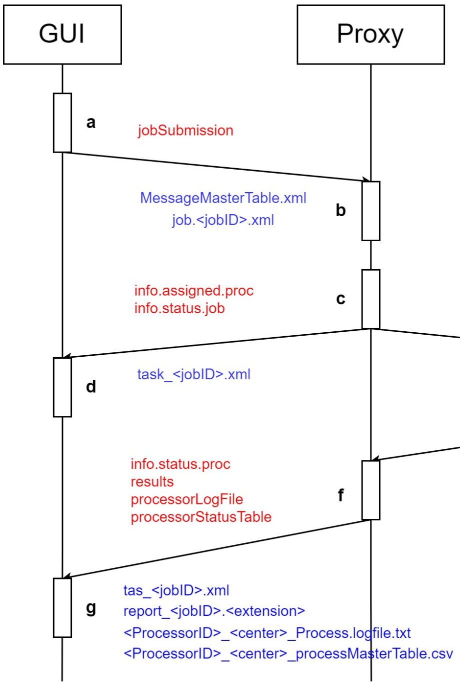


Figure 5: Communication flow that is triggered when a user submits a job from the GUI: **a.** The GUI updates the MessageMasterTable.xml, which maintains a comprehensive list of all submitted jobs, and generates a detailed file that describes the specific characteristics of each job. **b.** The Proxy constantly monitors the associated GUI URLs and reads the details from the *job_jobID.xml* file. **c.** The Proxy prepares the token (*token.zip*) for the Processor based on the information received from the GUI and deposits it in the designated processor's input folder for computation. **d.** The Proxy sends log messages to the GUI, providing information about the selected processor for task execution coded as *info.assigned.proc* and the job's status coded as *info.status.job*. **e.** The Processor triggers the task execution as soon as it finds the token.zip file in its input folder. **f.** The Proxy monitors the output and sync folders of the Processor and retrieves and forwards it to the GUI. The GUI processes the messages received from the Proxy regarding the processor's status.

3 Experimental setting

A GEN-RWD Sandbox instance was installed at the Fondazione Policlinico Universitario Agostino Gemelli, specifically in the Gemelli Generator - Real World Data facility's premises. This deployment aims to assess the usability of the GEN-RWD Sandbox for research purposes in the field of personalized medicine.

3.1 Methods

We tested our methodology on a real-world longitudinal clinical dataset of patients affected by Amyotrophic Lateral Sclerosis (ALS). Data were extracted from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) dataset [14], which comprehends demographic and clinical longitudinal information of patients enrolled in 23 distinct ALS clinical trials, and homogenized by selecting the patients with a shared panel of exams and the variables with less than 50% of missing values. Then, we further filtered out the visits with unknown time of onset, without a functional assessment, performed before the trial start, or with incomplete variables. This preprocessing resulted in 17995 records, each composed of 42 features both static (e.g., sex, age at onset, site of onset) and dynamically collected over the follow-up (e.g., the results of laboratory tests), referred to 1689 subjects. Data was subjected to K-means clustering as it is commonly used in medical research to find patients' phenotypes [15]. The selection of the number of clusters k was determined by employing the elbow method.

3.2 Results

To initiate the analysis, we utilized the GUI that allowed us to load the PRO-ACT dataset and select the desired clustering algorithm. Through the GUI, we also selected the laboratory variables on which we intended to train the model. The clustering algorithm identified three clusters within the dataset, with cluster number 2 having the highest level of compactness (Figure 6).

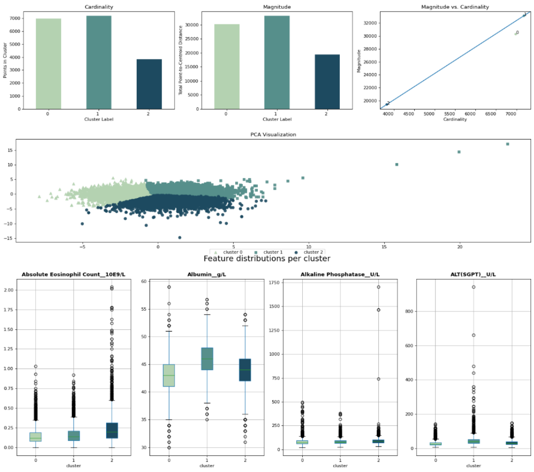


Figure 6: GEN-RWD Sandbox-generated report of K-means clustering trained over PRO-ACT data.

4 Discussion

This research proposes a modular architecture for developing a Distributed Analytics platform that is user-friendly and accessible to non-technical users. The architecture consists of three core modules that can be interconnected to serve different use cases. The ability to connect multiple graphical user interfaces (GUIs) to the same proxy allows for the creation of customized GUIs for different user types. Additionally, the Processor module can function independently to fulfill additional use cases, such as automating data processing tasks. Wirth et al. propose an evaluation framework for DA infrastructures in the context of medical research, focusing on both security and usability. The GEN-RWD Sandbox ensures data security by adhering to the principle of *Safe Data* within the Five Safes framework [17]. Data privacy is maintained by sharing only aggregated data, which prevents patient re-identification. This guarantees according to [16] also *Safe Outputs* and *Safe Settings*. The GEN-RWD Sandbox is a modular design that is scalable and flexible due to its modular design. It supports parallel processing and allows the execution of algorithms in any language as long as they are containerized within a Docker image. Compared to others, our solution offers a rather simple modular architecture, but it stands out as it is specifically designed for non-experts in data analysis, like doctors or clinical researchers. Future works on the platform include enhancing the GUI to further facilitate user analysis queries, expanding the Proxy module to make it compatible with authentication providers such as LDAP, and developing a system to connect the platform with heterogeneous data sources and perform data ingestion subject to preliminary data quality checks. Finally, the computational pipeline, from dataset selection to variable selection to algorithm execution, could be certified via blockchain.

Conflict of interests

The authors declare that they don't have conflicts of interests.

References

[1] A. M. Sebastian and D. Peter, "Artificial intelligence in cancer research: Trends, challenges and future directions," *Life*, vol. 12, no. 12, 2022.

[2] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council." [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[3] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, Mar. 2021.

[4] R. Hamza, A. Hassan, A. Ali, M. B. Bashir, S. M. Alqhtani, T. M. Tawfeeg, and A. Yousif, "Towards secure big data analysis via fully homomorphic encryption algorithms," *Entropy*, vol. 24, no. 4, 2022.

[5] N. Volgushev, M. Schwarzkopf, B. Getchell, M. Varia, A. Lapets, and A. Bestavros, "Conclave: Secure multi-party computation on big data," in *Proceedings of the Fourteenth EuroSys Conference 2019*, ser. EuroSys '19. New York, NY, USA: Association for Computing Machinery, 2019.

[6] G. M. Weber, S. N. Murphy, A. J. McMurry, D. MacFadden, D. J. Nigrin, S. Churchill, and I. S. Kohane, "The shared health research information network (shrine): a prototype federated query tool for clinical data repositories," *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 624–630, 2009.

[7] E. Jones, N. Sheehan, N. Masca, S. Wallace, M. Murtagh, and P. Burton, "Datashield–shared individual-level analysis without sharing the data: a biostatistical perspective," *Norsk epidemiology,sharing sensitive personal information. To cope with the systems of rules*, vol. 21, no. 2, 2012.

[8] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek *et al.*, "Observational health data sciences and informatics (ohdsi): opportunities for observational researchers," in *MEDINFO 2015: eHealth-enabled Health.* IOS Press, 2015, pp. 574–578.

[9] G. H. Lee, J. Park, J. Kim, Y. Kim, B. Choi, R. W. Park, S. Y. Rhee, and S.-Y. Shin, "Feasibility study of federated learning on the distributed research network of omop common data model," *Healthc Inform Res*, vol. 29, no. 2, pp. 168–173, 2023.

[10] "Clinerion ltd. patient network explorer solutions." [Online]. Available: https://www.clinerion.com/index/PatientNetworkExplorerSolutions.html

[11] U. Topaloglu and M. B. Palchuk, "Using a federated network of real-world data to optimize clinical trials operations," *JCO clinical cancer informatics*, vol. 2, pp. 1–10, 2018.

[12] O. Beyan, A. Choudhury, J. van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, M. R. Karim, M. Dumontier, S. Decker, L. O. B. da Silva Santos, and A. Dekker, "Distributed Analytics on Sensitive Medical Data: The Personal Health Train," *Data Intelligence*, vol. 2, no. 1-2, pp. 96–107, 01 2020.

[13] S. Welten, Y. Mou, L. Neumann, M. Jaberansary, Y. Yediel Ucer, T. Kirsten, S. Decker, and O. Beyan, "A privacy-preserving distributed analytics platform for health care data," *Methods Inf. Med.*, vol. 61, no. S 01, pp. e1–e11, Jun. 2022.

[14] N. Atassi, J. Berry, A. Shui, N. Zach, A. Sherman, E. Sinani, J. Walker, I. Katsovskiy, D. Schoenfeld, M. Cudkowicz *et al.*, "The PRO-ACT database design, initial analyses, and predictive features," *Neurology*, vol. 83, no. 19, pp. 1719–1725, 2014.

[15] N. Nidheesh, K. Abdul Nazeer, and P. Ameer, "An enhanced deterministic k-means clustering algorithm for cancer subtype prediction from gene expression data," *Computers in Biology and Medicine*, vol. 91, pp. 213–221, 2017.

[16] F. N. Wirth, T. Meurers, M. Johns, and F. Prasser, "Privacy-preserving data sharing infrastructures for medical research: systematization and comparison," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 242, Aug. 2021.

[17] L. Arbuckle and F. Ritchie, "The five safes of risk-based anonymization," *IEEE Security & Privacy*, vol. 17, no. 5, pp. 84–89, 2019.